MINING A DATABASE ON ALSATIAN RIVERS

C. GRAC, A. HERRMANN, F. LE BER
CEVH UMR MA 101, ENGEES, 1 quai Koch
B.P. 61039, 67070 Strasbourg, France

M. TRÉMOLIÈRES
CEVH UMR MA 101, Institut de Botanique, 28 rue Goethe
67083 Strasbourg cedex, France

A. BRAUD, A. HANDJA, N. LACHICHE
AFD, LSITT UMR 7005, Bd Sébastien Brant
BP 10413, 67412 Illkirch cedex, France

We aim at comparing the answers of the bio-indication tools of the water bodies towards the various pressures which they undergo. Therefore we have designed of a database gathering the physical, physico-chemical, floristic (diatoms and macrophytes), and faunistic data (invertebrates, oligochaetes, fishes) of the rivers streams and water areas of the Alsace plain. Besides, we are implementing data mining methods to explore the database. These methods are used to find out regularities and similarities that can be interpreted by a domain expert. Indeed, we consider data mining as a part of a more general process that is knowledge discovery in databases. We give examples of such methods and the results that can be obtained.

INTRODUCTION

Water quality is a major problem in Europe, underlined by the recent European Water Framework Directive. To evaluate the physico-chemical quality of a water body appeared to be not sufficient, and new tools are required for evaluating the quality of the whole ecosystem [7]. This implies to compare and combine the existing tools, such as biological indices, which focus only on a part of the ecosystem.

Our project aims at handling the comparison of biological indices in the Alsace plain. Actually, many hydrochemical and hydrobiological data were existing on the rivers streams and water areas of the Alsace plain. Therefore, the design of a database gathering the physical, physico-chemical, floristic (diatoms and macrophytes), and faunistic data (invertebrates, oligochaetes, fishes) appeared to be necessary. Indeed, in addition to structuring the data, this base will be used for comparing the answers of the bio-indication tools of the water bodies towards the various pressures which they undergo. Thus, the database includes the values of the French biological indices of running water quality, based on the five groups above, and the data related to the biological features

of these same groups. The data on the macrophytes, acquired for twenty years, will be supplemented by faunistic and other floristic (diatoms) data. According to this aim, sites were selected on the basis of the identifications of the water bodies in Alsace and the undergone pressures. These identifications were carried out in the Rhine-Meuse basin according to the European Water Framework Directive.

Besides, we are implementing data mining methods to explore the database. These methods are used to find out regularities and similarities that can be interpreted by a domain expert. Indeed, we consider data mining as a part of a more general process that is knowledge discovery in databases. Several methods are possible, according to the characteristics of the data, which are multi-relational, temporal, spatial, quantitative or qualitative, and incomplete. In particular, we use symbolic and numerical classification methods, and machine learning methods. Eventually, a knowledge base will be implemented to help the knowledge discovery process.

In this paper, we first present the general aim of our project and the structure of the database. In a second part, we describe a few data mining methods and the results that were obtained in related work on hydrobiological data. Then we conclude on our future work.

STUDYING WATER QUALITY

Comparing indices

In France, five biological indices have been normalized to assess quality of running water. They are based on three faunistic groups: the invertebrate index [1], the oligochaete index [2], the fish index [3], and on two floristic groups: the diatom index [4], and the macrophyte index [5].

According to AFNOR (French organism of normalization) [1,2,3,4,5] and [17], each of them estimates water quality differently. The macrophyte index estimates the trophic level of water, the diatom index gives the global water quality, the oligochaete index gives an evaluation of the sediment quality, and the fish index allows to classify the chemical and physical water quality quite like the invertebrate index. Therefore, their answers on a same station, with a same undergone pressure, at the same time can be really different but the simultaneous application of these five indices is not common and work comparing their answers are not frequent [17]. Furthermore, according to the French Ministry of Environment [7], in order to applicate the new European Water Framework Directive, the five indices need to be adapted and it is necessary to conceive new tools able to estimate the quality of global water ecosystem.

Lafont [17] proposed "*the ecological ambiance system*", a system associating the five indices based on the different biological groups and able to estimate the quality of the whole complex water ecosystem. Our objective is to try to develop this concept and to propose a tool concretly applicable. We rely therefore on a large database collecting data on alsatian rivers streams and water areas.

A database on alsatian rivers

Our data have been collected from the streams of Alsace plain hydroecological region [22]. They are environnemental data, such as the discharge, or the current weather; physical data, such as the existence of a structure; chemical data, such as nitrates, phosphates, organic matter; floristic data; and faunistic data. The floristic groups we have collected are the diatoms, and the macrophytes. The faunistic groups are the invertebrates, the oligochaetes, and the fishes.

A part of the data was collected earlier by our team or by the public organisms in charge of the national water monitoring. Another part is being collected. The oldest data were collected for 20 years, from around 700 stations [21]. These data are incomplete: for most stations, the environmental, physical and chemical data are present, but there are only one faunistic or floristic group. The recent data are complete, except for the fishes. They have been collected on 40 places in Alsace , that have been chosen among the seven identified types of water bodies in this hydroecological region. For each type, a reference station —with no anthropic pressure— and other stations with various pressures have been selected. The sampling methods used for the faunistic and floristic data are the normalized methods of the french biological indices, modified according to the recommendations of the european research program AQEM [6].

In order to keep all these data, we have developed a specific database system [14]. This database was built according to the french national format « SANDRE » for water data. The database consists of 34 tables. The main path corresponds to a succession of tables from the river to the hydrochemical and hydrobiological data: sites visited, sampling dates, environmental conditions, sampling methods used, chemical results, biological results (Fig. 1).

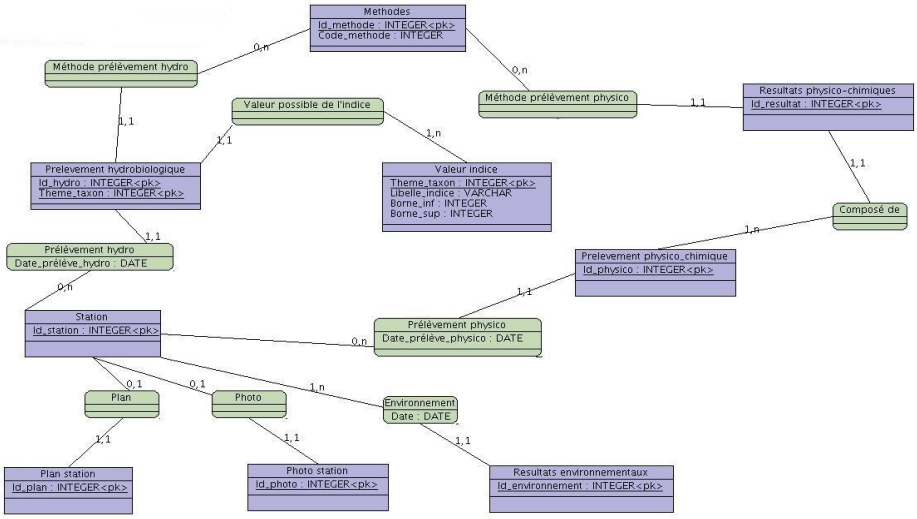

Figure 1. Part of the database tables: the station, sampling methods and results tables, with their links.

The other tables contain informations about the different parameters, essentially chemical and bological parameters. For instance, each floristic or faunistic taxon, with its taxonomic nomenclature, is in a specific table called « Taxon ». This table comes from SANDRE (about 2000 records) and has been completed (about 4500 records). Three tables linked to the table "Taxon" detail the characteristics of each taxon.

The database structure was implemented within the MySQL language and was given a web interface to be used at distant sites. The user interface allows to add and to extract the main data [14].

MINING RIVER DATA

Data mining offers an interesting range of techniques for exploring large databases and helping data interpretation [9,15]. Work in the area of river data varies with respect to the task tackled and the technique used. They often aim at designing models allowing a better understanding of the relations between the diversity of a small community of living organisms and the biological, environmental or physico-chemical characteristics of the water body. Let us notice that systems used to evaluate water quality are different accross countries, which also results in differences in the research works, people focussing

on different parameters. Finally, the expert knwoledge, and his/her expectations, have a major part in these studies.

Related Work

In his thesis, Goethals [16] dealt with the problem of predicting macroinvertebrates communities present in rivers using data mining techniques. He reported approaches based on fuzzy logic, bayesian belief networks, artificial neural networks (ANNs), classification and regression trees. His own work focussed on two of those techniques, namely ANNs and classification trees.

ANNs have been widely studied in ecological modelling as the ecological systems are highly nonlinear. However, the fact that they give good prediction results may be weakened by the black-box character of the approach which makes it difficult to understand [8]. This is a severe drawback for what concerns the validation by the expert and thus it prevents from tuning the system according to the remarks the expert may formulate on the results. From that point of view, symbolic models, such as classification and regression trees, represent a good alternative [10].

Fewer work exists in that area. In [13], the authors worked on data related to British and Slovenian rivers. They dealt with the tasks of predicting a class of abundance for taxa using the CN2 system for rule induction. CN2 learns IF-THEN rules, that enable to describe the conditions under which a property will be fulfilled. For example, the rule IF *pH=7* THEN *water_quality=good* means that if the pH of the water is equal to 7 then the water quality is good. The general form of these rules is IF *set_of_conditions* THEN *conclusions*. In the case of CN2, conditions are of the form "*attribute comparator value*". In this paper, several tasks were addressed: predicting the water quality based on biological or physico-chemical properties, and analysing the influence of physico-chemical properties on the presence of taxa. The readibility of the rules allowed interactions with the expert, so that the authors were able to tune the system by adding new attributes capturing information used by hydrobiologists in their evaluations. The obtained rules contained valuable knowledge, sometimes complementing the knowledge of the expert. As expected, physico-chemical properties did not fully determine the presence of taxa as they change quickly, while living organisms are affected on the long term by the pollutions they induce.

Further researches were carried out by the authors on physico-chemical parameters in [11]. Regression trees were used in order to predict physico-chemical parameters based on biological ones. Regression trees are also easily understandable. Moreover, their structure shows the parameters which have the strongest weight in the

prediction, as they are the closest to the root of the tree. Two kinds of studies were made, one where taxa were considered at the level of species and the other one where they were considered at family level. Results at the species level were better, which was expected since it is more precise. However, the structure of one of the trees learnt at the family level revealed the importance of a particular family for evaluating the water quality. It is interesting to note that this family did not appear as so important when inspecting the trees learnt at the species level and that this fact had been suprising for the expert. The combination of both approaches was thus investigated but did not result in improvements.

In [12], regression trees were used to investigate further the relations between physico-chemical properties of water and the diversity of living organisms. The relations were studied by trying to predict the number of taxa based on physico-chemical properties. This was extended in [8], where several physico-chemical parameters were predicted at the same time. This way, one could see the interdependence of parameters. The overall predictive accuracy was the same for the parameters in both cases, but for some it was better if computed at the same time as other parameters. A second part of the paper investigated the prediction of past physico-chemical properties from biological ones, represented by the minimum, maximum and average values computed on the three months preceding the date of the biological sample. This gave insights on the way physico-chemical properties may influence biological ones. Results confirmed the expectations of the expert.

These works show how symbolic methods facilitate the discussion with experts, and how it can help in tuning the system. However, questions remains open on the relations between the parameters. Moreover, if possible, more properties, such as environmental ones, should be taken into account.

Mining a database on alsatian rivers

According to the data mining studies described above, which show the interest of symbolic approaches in the area of hydrobiological data, we will experiment those approaches on our data. Firstly we try to repeat the results obtained in [8,11,12], concerning the prediction of physico-chemical properties from biological ones or the prediction of taxa based on physico-chemical properties. Such results can be useful to complement our database, especially for the ancient data where some parameters are lacking. This work is in progress.

Then, we will explore rule induction methods, such as concept lattices [18], or inductive logic programming [19]. These approaches involve the design of a model of the domain knowledge, or *ontology*. This ontology, based on the expertise of some members of CEVH will be linked to the existing database in order to build a general k*nowledge*

discovery system. It could be implemented within a descriptions logic language, such as OWL [20], taking advantage of the open source systems used for the implementation of the database.

CONCLUSION

The general aim of our work is to try to propose a floristic, faunistic, "indicial" typology by water body and the undergone pressure. This work takes part in the reflexion carried out on the follow-up means of the ecological state of the rivers streams and the achievement of their good state. To achieve our aim, we rely on a set of data on alsatian rivers, and the expertise of CEVH members in hydrobiology. Furthermore, we have built a specific database system and we are experimenting data mining techniques to help the analysis of this database. First results will be presented at the conference.

REFERENCES

[1] AFNOR, "Qualité de l'eau : détermination de l'Indice Biologique Global Normalisé (IBGN)", *NF T90-350* (1992).

[2] AFNOR, "Qualité de l'eau : détermination de l'Indice Oligochètes de Bioindication des Sédiments (IOBS)", *NF T90-390* (2002).

[3] AFNOR, "Qualité de l'eau : détermination de l'Indice poissons rivière (IPR)", *NF T90-344* (2004).

[4] AFNOR, "Qualité de l'eau : détermination de l'Indice Biologique Diatomées (IBD)", *NF T90-354* (2000).

[5] AFNOR, "Qualité de l'eau : détermination de l'Indice Biologique Macrophytique en Rivière (IBMR)", *NF T90-395* (2003).

[6] AQEM (developement and testing of an integrated Assessment system for the ecological Quality of streamsand river throughout Europe using benthic Macroinvertebrates), http://www.aqem.de

[7] Bazerques M.-F., "Directive-cadre sur l'eau : le bon état écologique des eaux douces de surface : sa définition, son évaluation". *Communication au Ministère de l'Ecologie et du Développement Durable* (2004).

[8] Blockeel H., Džeroski S., and Grbović J., "Simultaneous prediction of multiple chemical parameters of river water quality with TILDE", *Proc. Third European Conference on Principles of Data Mining and Knowledge Discovery*, (1999), pp 15-18.

[9] Dunham M. H., "Data Mining. Introductory and Advanced Topics", Prentice Hall (2003).

[10] Džeroski S., "Applications of symbolic machine learning to ecological modelling", *Ecological Modelling*, Vol. 146, (2001), pp 263-273.

[11]Džeroski S., Demšar D. and Grbović J., “Predicting chemical parameters of river water quality from bioindicator data”, *Applied Intelligence*, Vol 13, No 1, (2000), pp 7-17.
[12]S. Džeroski, and Grbović J., “Relating biodiversity of river communities to physical and chemical water properties”, *Proc. Sustainability in the information society (15th Int. Symposium on Informatics for Environmental Protection)*, Marburg, Metropolis, Part. 1, (2001), pp 367-372.
[13]Džeroski S,. Grbović J., Walley W. J., and Kompare B., “Using machine learning techniques in the construction of models. Part II: Rule induction”, *Ecological Modelling*, Vol. 95, (1997), pp 95-111.
[14]Ehrhard J.-L., “Mise en œuvre d'un système de comparaison des réponses des indices biologiques sur les cours d'eau de la plaine d'Alsace”, Mémoire d'ingénieur CNAM, Strasbourg (2005).
[15]Fayard U., Piatesky-Shapiro G., Smyth P. and Uthurusamy R., “Advances in Knowledge Discovery in Data Mining”, AAAI Press / The Press (1996).
[16]Goethals P., “Data driven development of predictive ecological models for benthic macroinvertebrates in rivers”, PhD Thesis, Universiteit Gent. Faculteit Bio-ingenieurswetenschappen (2005).
[17]Lafont M., "A conceptual approach to the biomonitoring of freshwater: the ecological ambience system", *Freshwater J. Limno.,* Vol. 60 (supp.1), (2001), pp 17-24.
[18]Napoli A., “A smooth introduction to symbolic methods for knowledge discovery”. Research report, LORIA, Nancy (2005).
[19]De Raedt L., Blockeel H.,. Dehaspe L and Van Laer W., “Three Companions for Data Mining in First Order Logic”, *Relational Data Mining,* pp 105-139, Springer-Verlag (2001).
[20]Smith M.K., Welty C., and McGuinness D.L. (Editors), OWL Web Ontology Language Guide, W3C Recommendation (2004).
[21]Trémolières M., Carbierner R., Orstcheit A., and Klein J.-P., "Changes in aquatic vegetation in Rhine floodplain streams in Alsace in relation to disturbance", *J. Veget. Sc.,* Vol. 5, (1994), pp 169-178.
[22]Wasson J.G., Chandesris A., Pella H., and Blanc L., "Les hydroécorégions de France métropolitaine - approche régionale de la typologie des eaux courantes et éléments pour la définition des peuplements de référence d'invertébrés", CEMAGREF (2002).